Optimization Methods in Science and Engineering

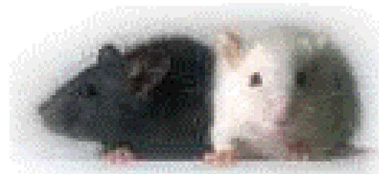
Juan Meza

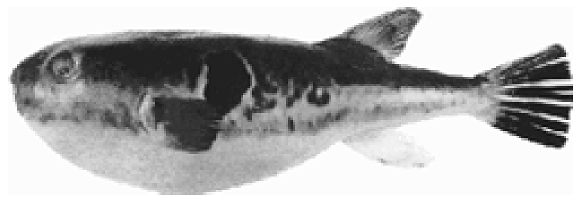
High Performance Computing Research
Lawrence Berkeley National Laboratory
http://crd.lbl.gov/~meza

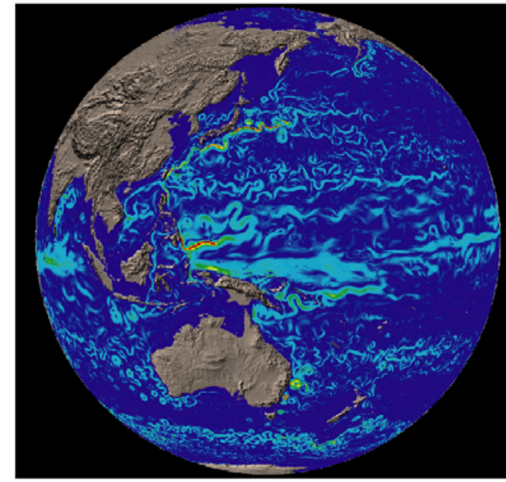
Lawrence Berkeley National Laboratory

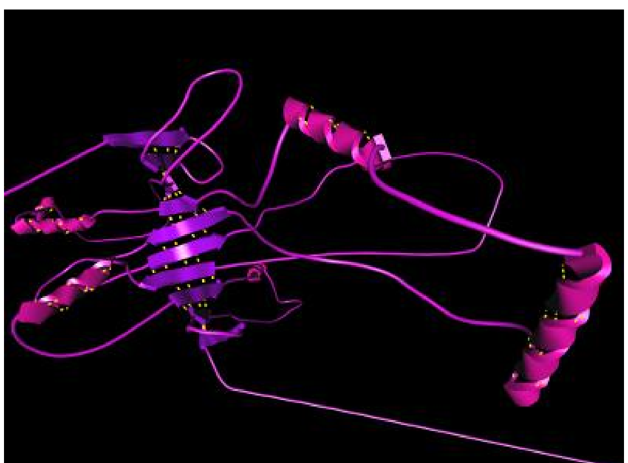
- Department of Energy national laboratory
- Open, unclassified, basic research
- Home to NERSC, the fifth largest supercomputing center in the world (7.3 Tflops)
- Located in the hills next to University of California, Berkeley campus

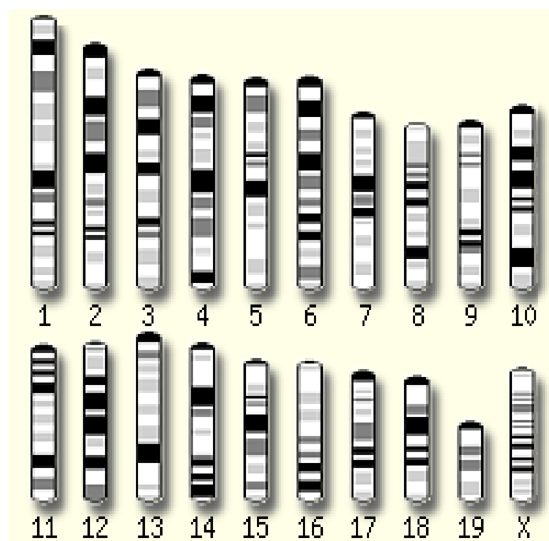
LBNL sponsors a wide range of computational sciences activities

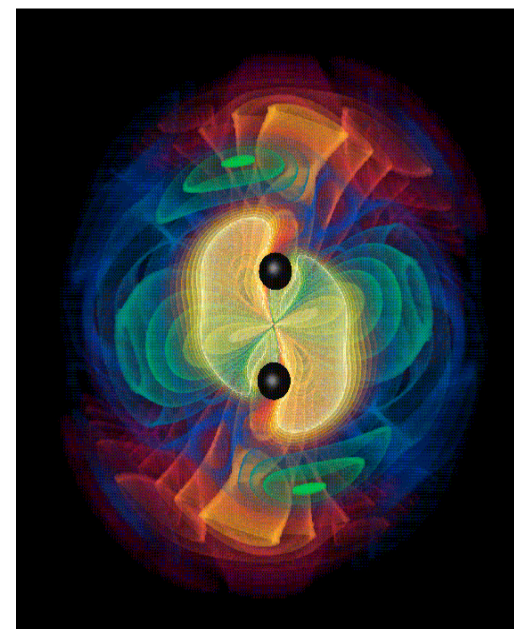




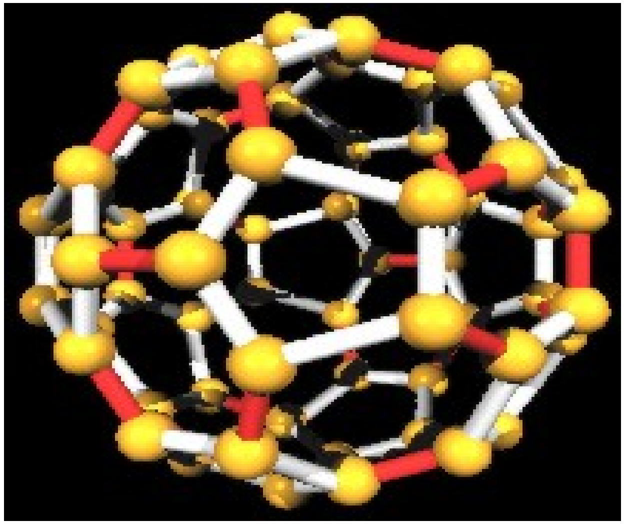


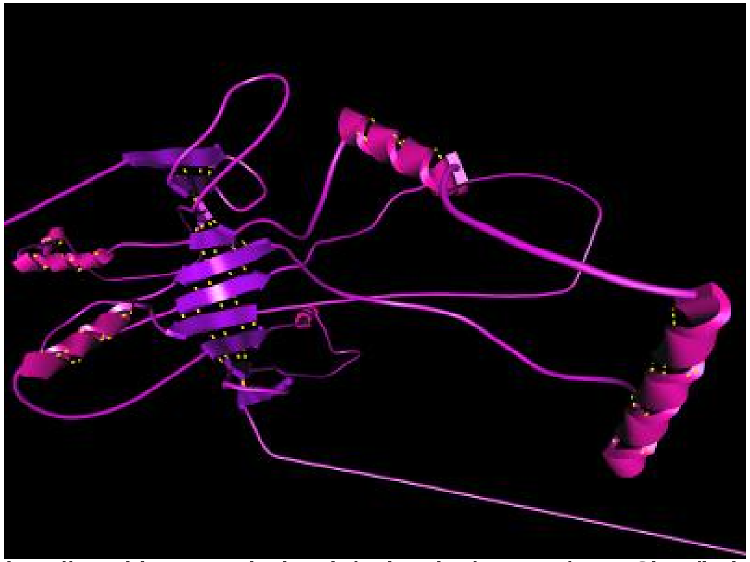






Modeling and simulation often involves optimization

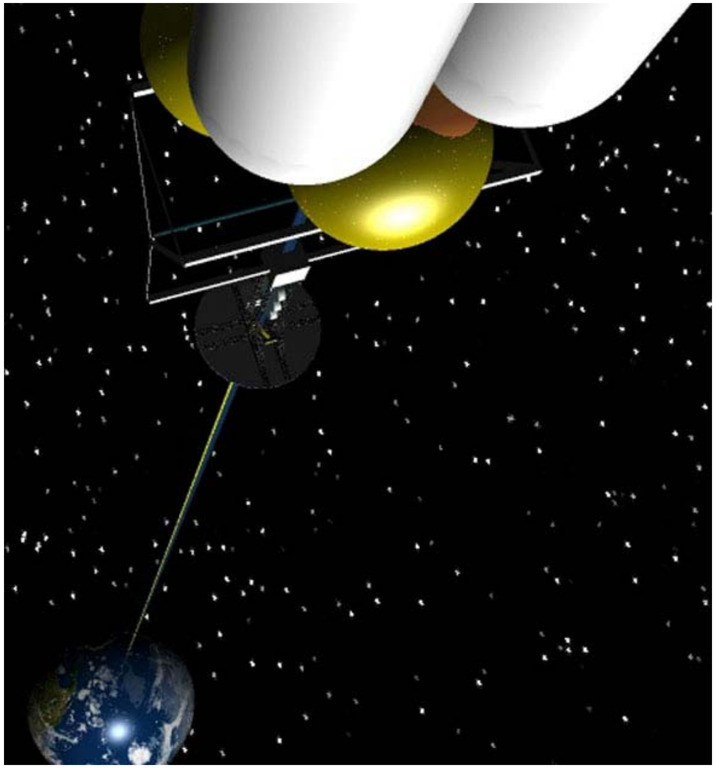


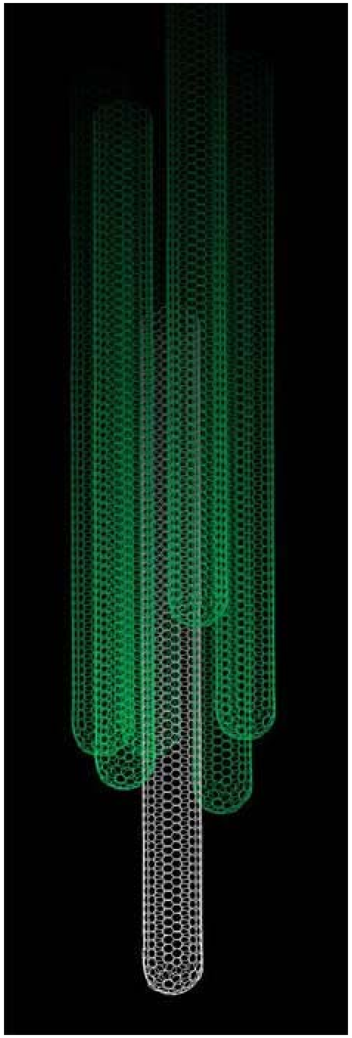


http://graphics.cs.ucdavis.edu/~okreylos/ResDev/ProtoShop/index.html

- Predict properties of nanostructures or design nanostructures with desired properties
- Protein folding problems attempt to construct 3D structures from a linear sequence (the genome)
- These simulation-based optimization problems have different characteristics than standard problems

World's tallest elevator!

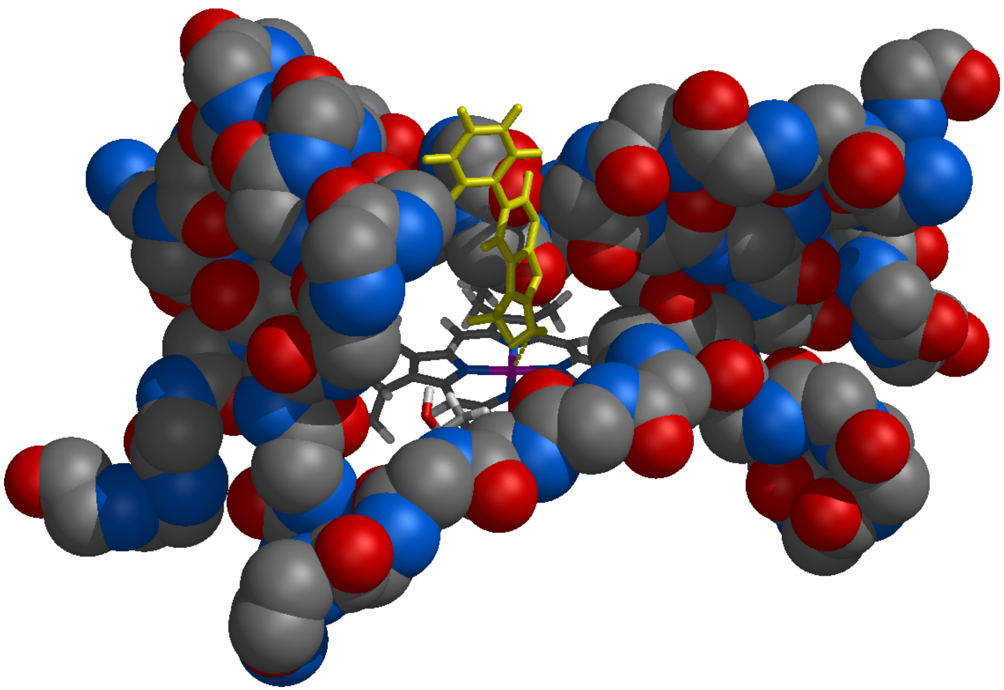




1) NY times, Sept. 23, 2003.
2)Tech Wednesday, March 2002

- Idea is to build an elevator 60,000 miles high to carry cargo into space
- Concept is based on designing ultrastrong fiber strands from carbon nanotubes
- These ribbons of nanotubes would be woven into one paper-thin meter-wide ribbon

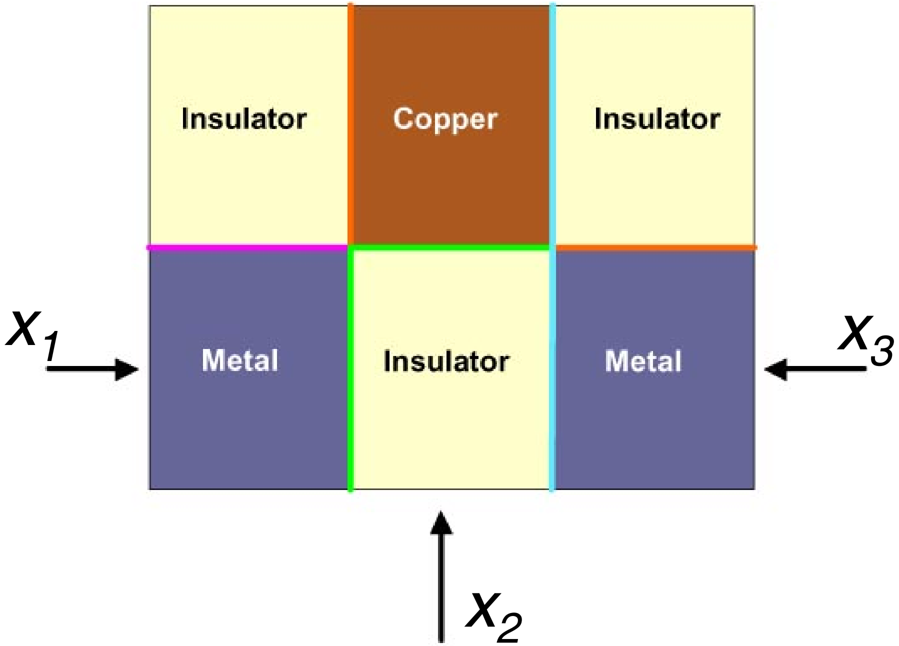
Molecular structure prediction

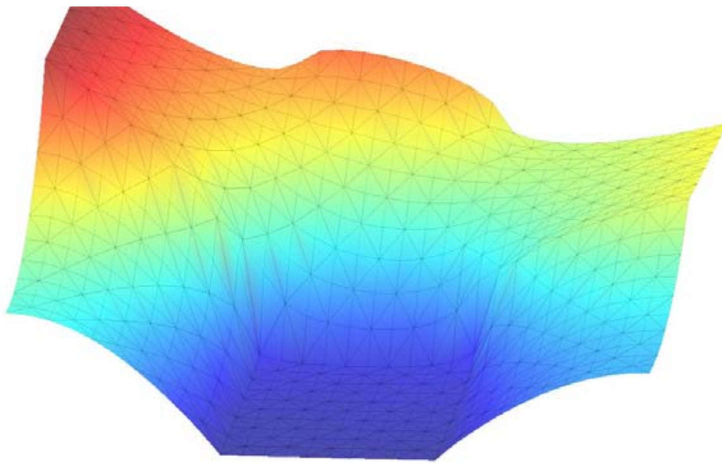


Docking model for environmental carcinogen bound in *Pseudomonas Putida* cytochrome P450

- A single new drug may cost over $500 million to develop and the design process typically takes more than 10 years
- There are thousands of parameters and constraints
- There are thousands of local minima

Parameter identification

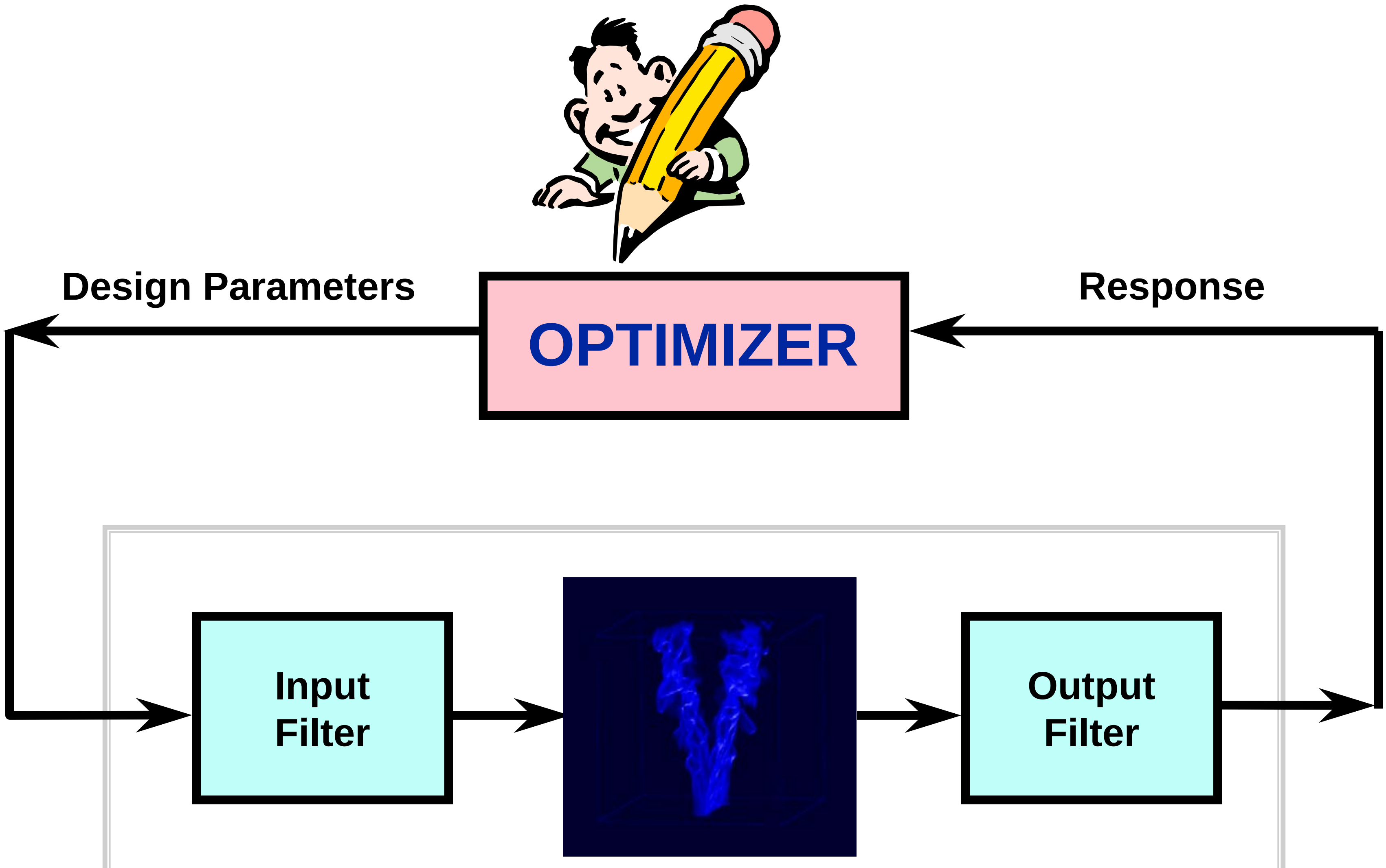




- Find model parameters, (satisfying some bounds), for which the simulation matches the observed temperature profiles
- Objective function consists of computing the temperature difference between simulation results and experimental data:

$$\min_{x} \sum_{i=1}^{N} (T_i(x) - T_i^*)^2$$

Optimization can be used in conjunction with simulation codes



General Optimization Problem

$$\min_{x \in \Re^n} f(x),$$ Objective function

$$s.t. \quad h(x) = 0,$$ Equality constraints

$$g(x) \geq 0$$ Inequality constraints

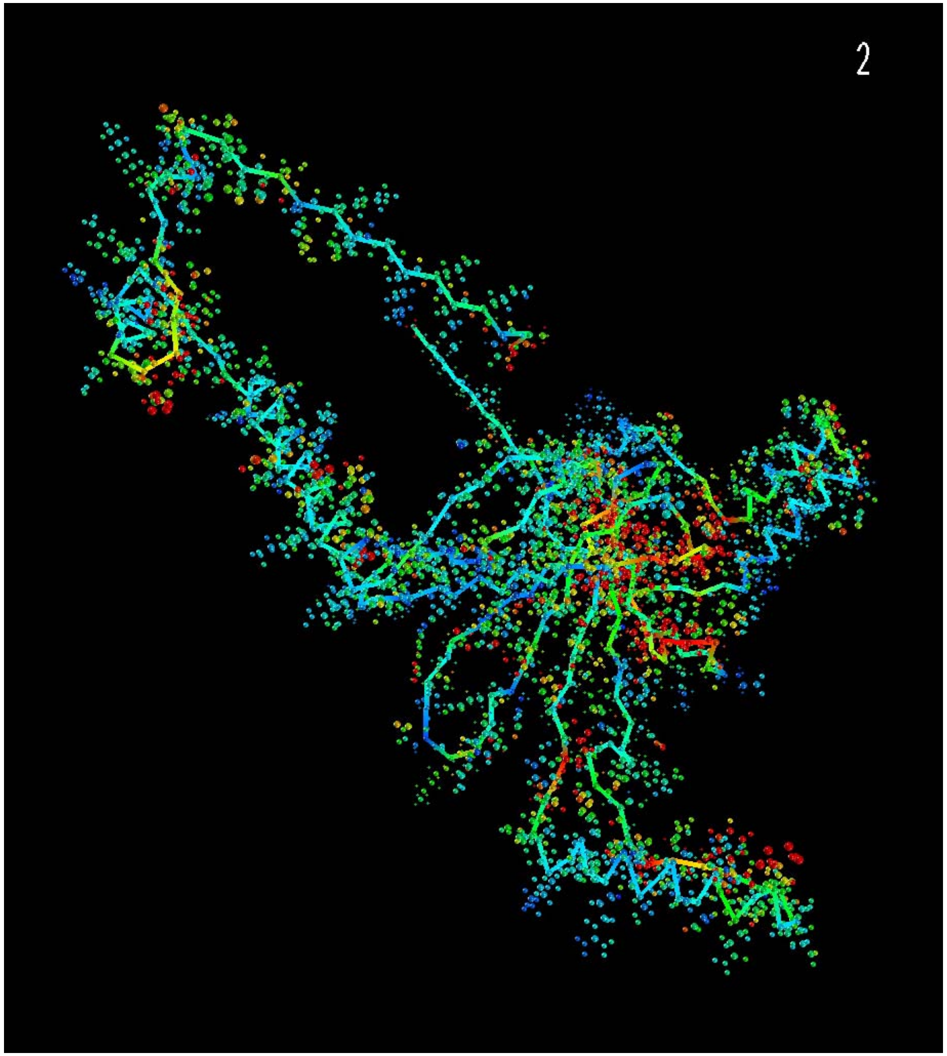
Optimization Problem Types

- Unconstrained optimization
- Bound constrained optimization
  - Only upper and lower bounds
  - Sometimes called “box” constraints
- General nonlinearly constrained optimization
  - Equality and inequality constraints
  - Usually nonlinear
- Some special case classes
  - Linear programming (function and constraints linear)
  - Quadratic programming (quadratic function, linear constraints)

Why are simulation-based optimization problems different?

- Objective function is smooth
  - Usually true, but simulations can create noisy behavior
- Twice continuously differentiable
  - Usually true, but difficult to prove
- Constraints are linearly independent or hard
  - Users can sometimes over-specify or incorrectly guess constraints
  - Require strict feasibility
- Expensive objective functions
  - Dominant cost is evaluation of function

Energy Minimization Using Limited Memory BFGS (LBFGS)



Protein T162 (from CASP5)

- Energy Function: AMBER
- Protein 162;
- N = 13728 (4576 Atoms)
- LBFGS with M=15
- Total number of LBFGS iterations = 11656
- Total number of function evaluations = 11887
- Each function evaluation takes approximately 5 CPU sec

Amber Function

$$E_{AMBER} = E_{Bonds} + E_{Angles} + E_{Dihedrals} + E_{NonBonded}$$

$$E_{Bonds} = \sum_{\text{Bonds}} K_{r_i} (r_i - \bar{r}_i)^2$$

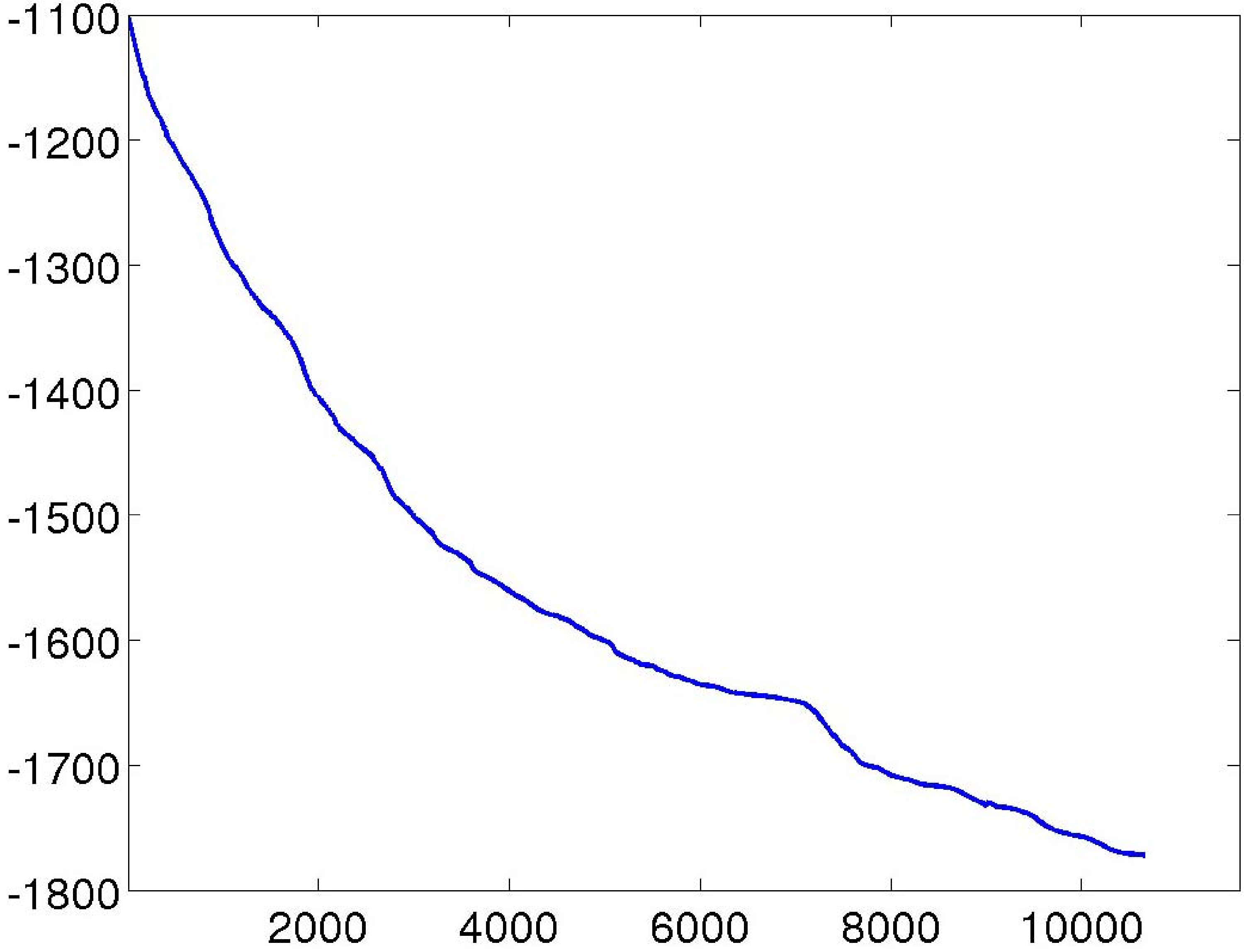
$$E_{Angles} = \sum_{\text{Angles}} K_{\theta_i} \left(\theta_i - \bar{\theta}_i\right)^2$$

$$E_{Dihedrals} = \sum_{\text{Dihedrals}} K_{\phi_i} \left(1 + \cos(n_i \phi_i - \delta_i)\right)$$

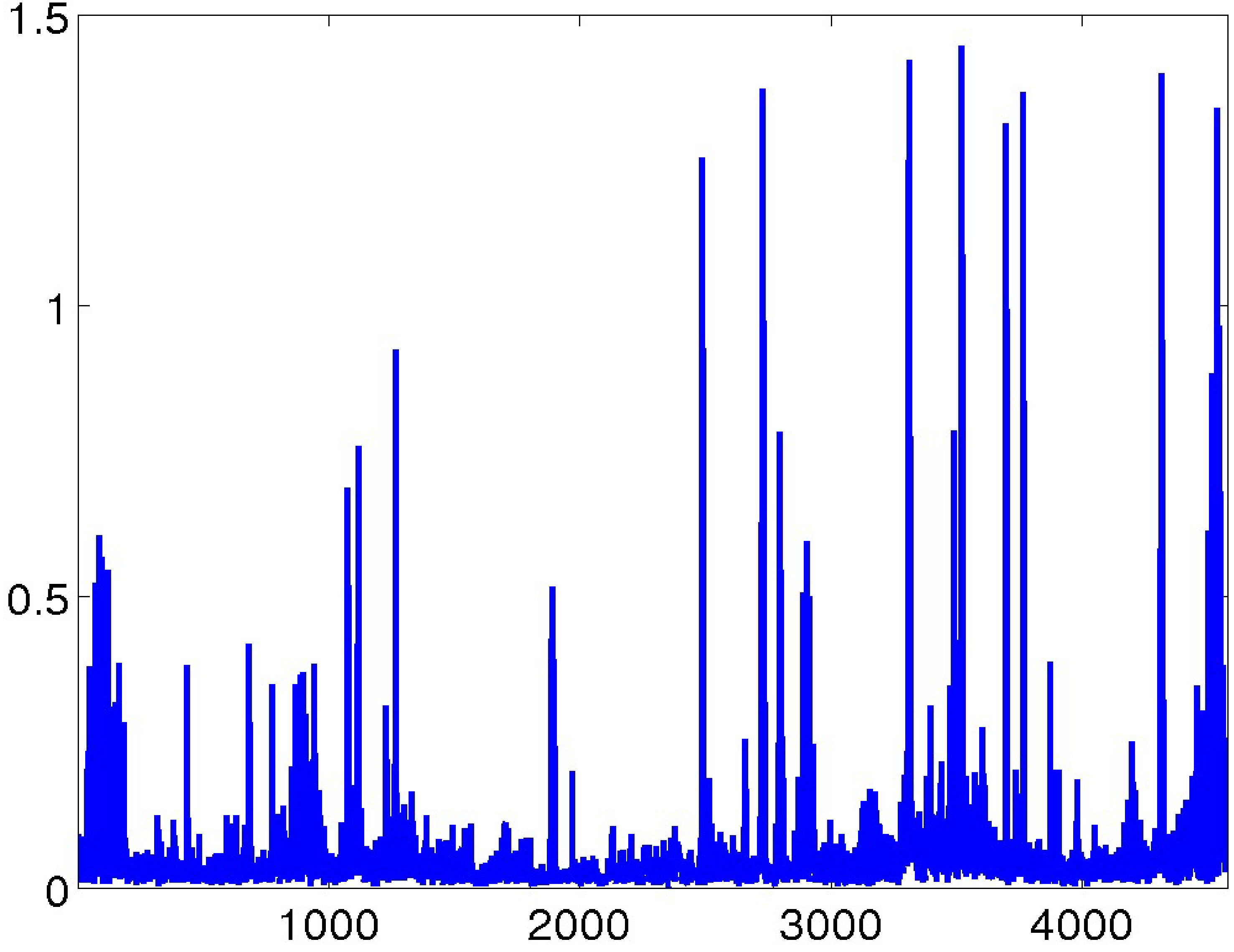
$$E_{NonBonded} = \sum_i \sum_{i<j} \left(\varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{r_{ij}} \right)$$

A Physical Approach to Protein Structure Prediction, Crivelli, et.al. Biophysical Journal, Vol 82, 2002.

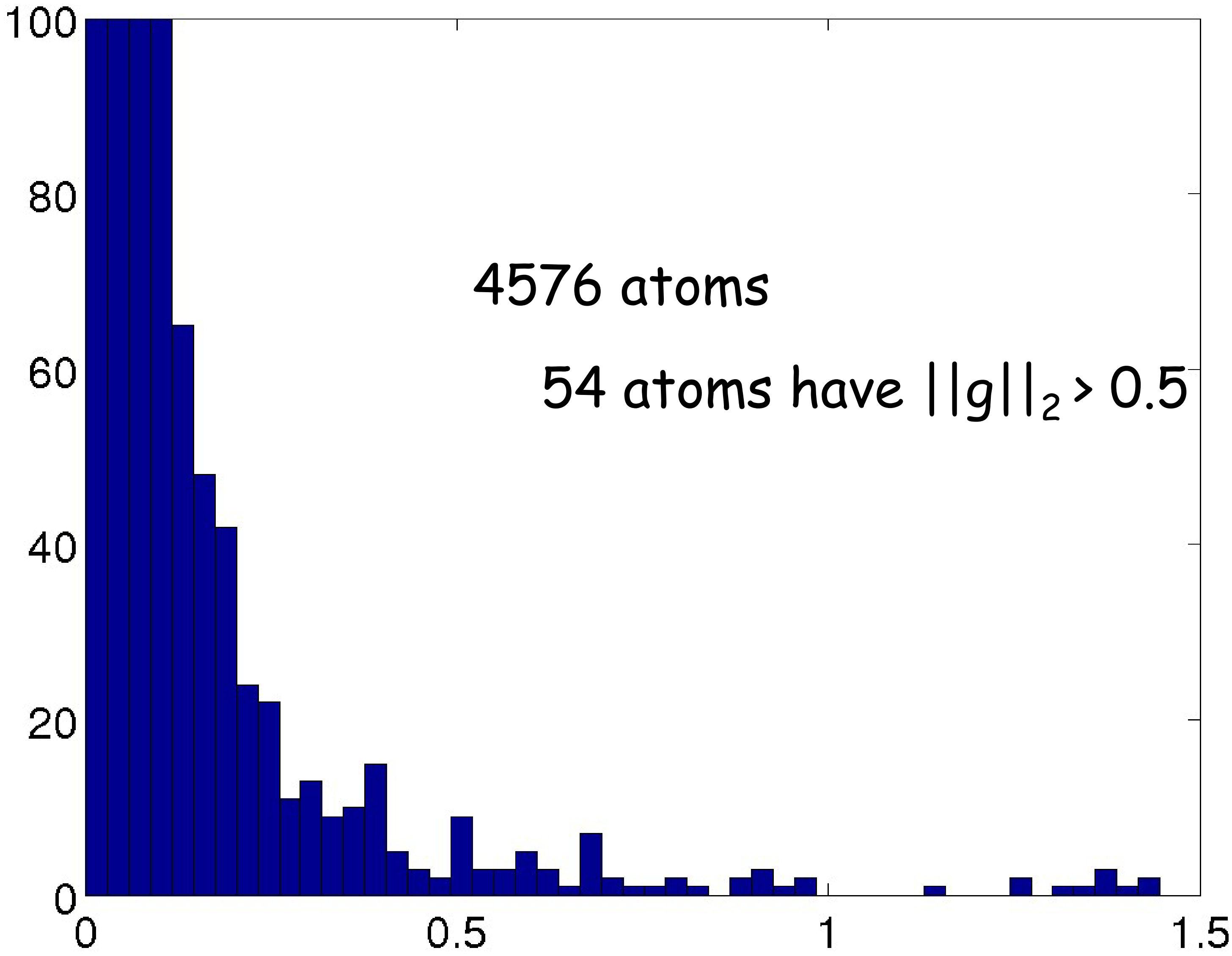
Energy vs. LBFGS iterations for T162 Problem



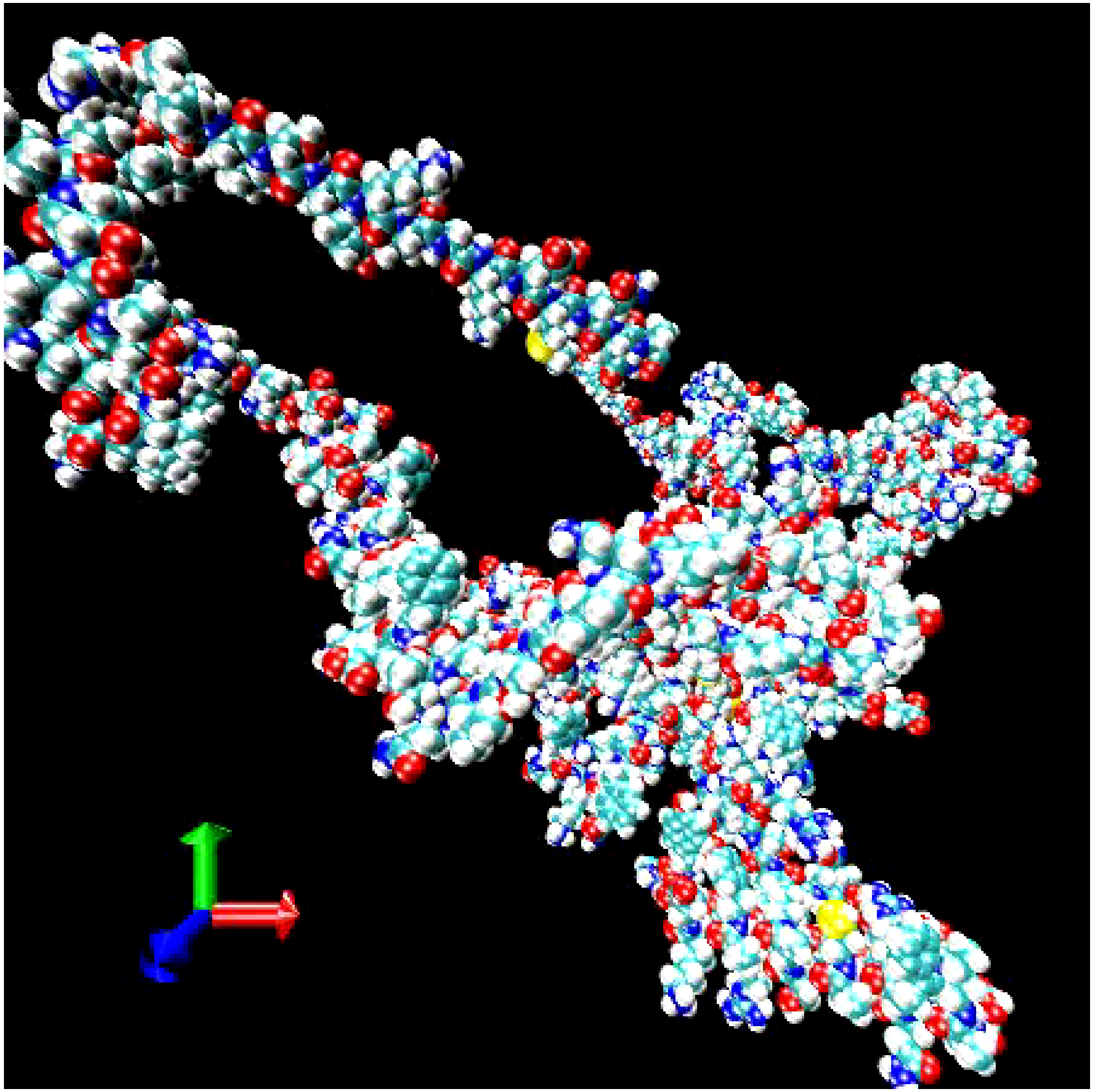
T162 Protein: ||gradient|| by atom



Distribution of ||gradient|| by atom



Protein T162 (from CASP5)



- Initial configuration created using ProteinShop (S. Crivelli)
- Energy minimization computed using OPT++/LBFGS
- Final average RMSD change was 3.9 Å
- Total simulation took approximately 32 hours on a 1.7GHz machine

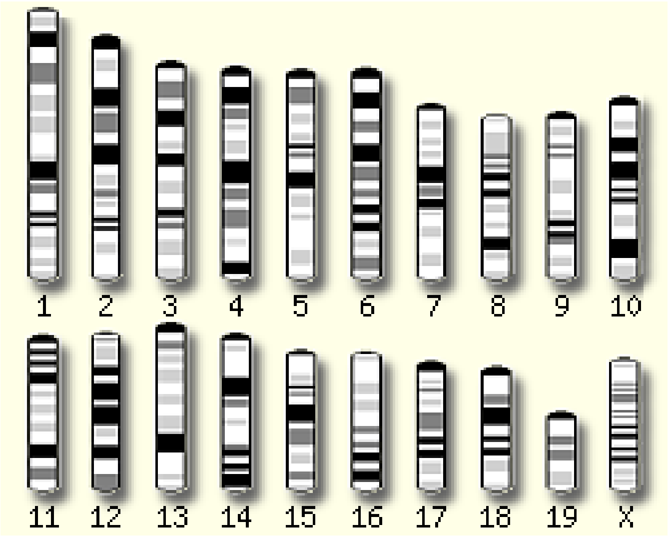
Summary

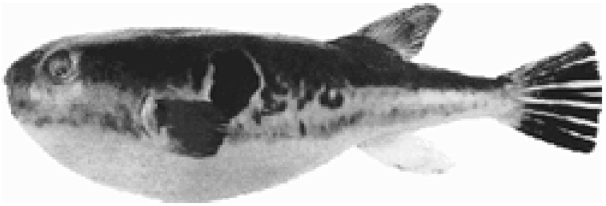
- Wide range of scientific and engineering problems requiring mathematics
- Many of these scientific and engineering problems involve nonlinear optimization problems
- Thorough knowledge of both science and mathematics is required to address these problems - the solution of these problems requires interdisciplinary teams, creativity, and a little bit of luck.

Questions ?

Backup Slides

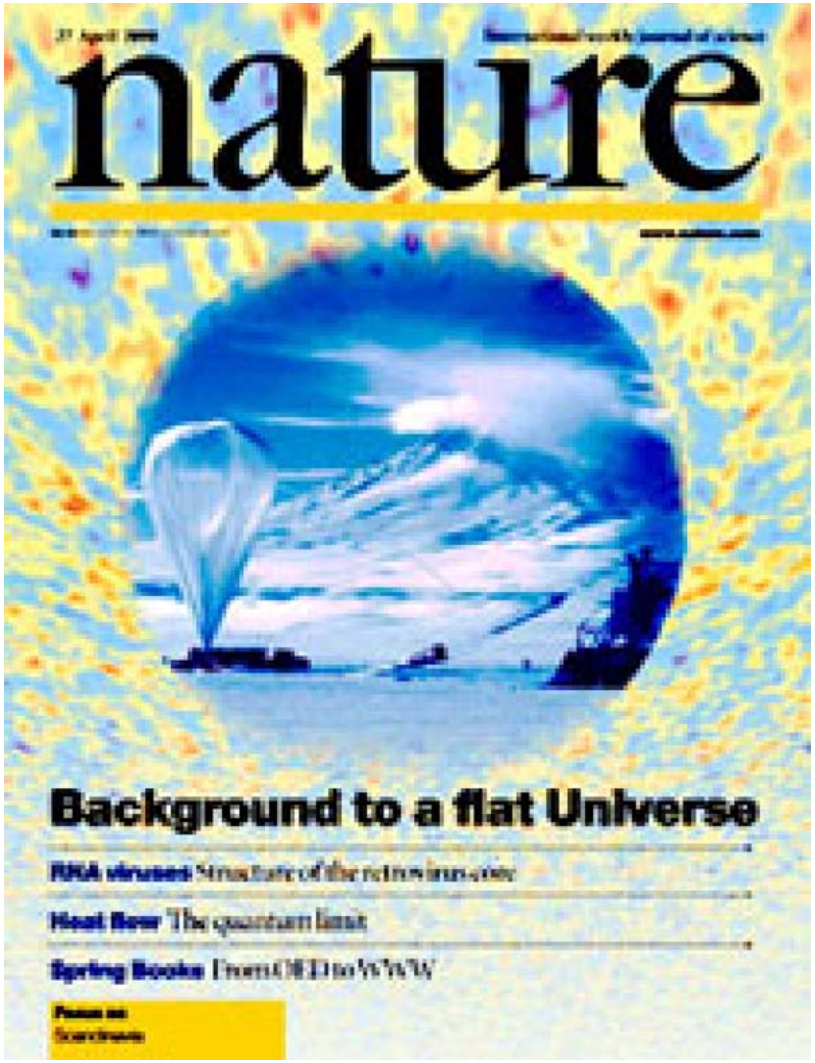
JAZZ Genome Assembler





- Assembly of Fugu genome from 3.1 million reads, and initial preparation of mouse genome data.
- NERSC provided:
  - porting of JAZZ assembler, BLAST alignment tool, cross_match alignment tool, and MySQL client to the IBM SP
  - a dedicated MySQL server
  - resolved issues installing a MySQL server on the IBM SP
  - consulting support for parallelization of BLAST and cross_match tool
- Dan Rokhsar, Joint Genome Institute

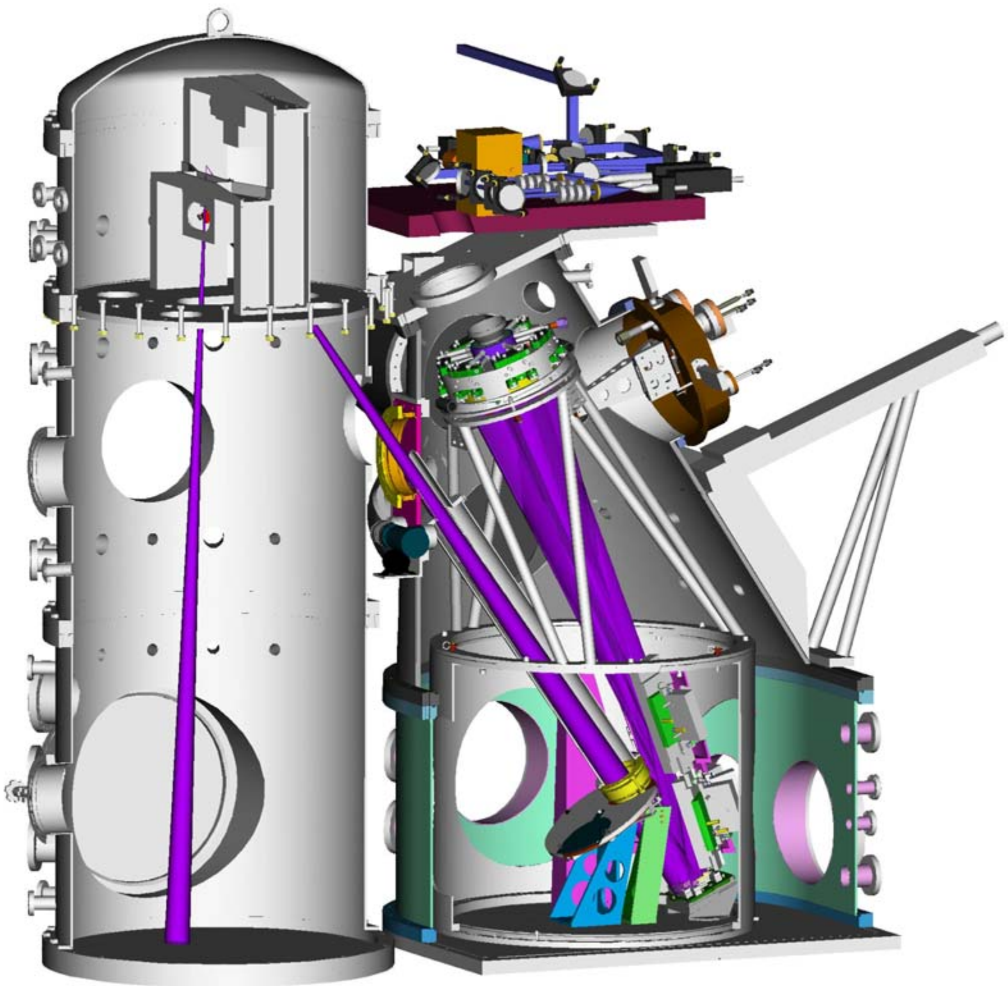
Analyzing Cosmic Microwave Background Radiation



- BOOMERANG Experiments – analyze cosmic microwave background radiation data to obtain a better understanding of the universe
- The data analysis provides strong evidence that the geometry of the universe is flat
- Computational capability provided on NERSC platforms
- MADCAP software developed at NERSC for general community

Borrill (LBNL) + CalTech + others.
April 27, 2000

Parameter identification example



- Find model parameters, satisfying some bounds, for which the simulation matches the observed temperature profiles
- Computing objective function requires running thermal analysis code
- Each simulation requires approximately 7 hours on 1 processor